1. Objasniti pojam dokumenta, papirnog dokumenta i digitalnog dokumenta?
Pojam dokumenta obuhvata:
* tradicionalne papirne dokumente
* računraski dorađene informacije kojima se rukuje kao osnovnom jedinicom obrade

2. Šta su to metapodaci i navesti nekoliko primera metapodataka?
Metapodaci su podaci o podacima. Sadržaj elektronskog dokumenta predstavlja podatke. Podaci o dokumentu predstavljaju metapodatke. Postoje li metapodaci za metapodatke? formalno mogu da postoje pitanje je koliko to ima smisla.
Primeri:
metapodaci za teksturalni dokument: #autor, #naslov, #datumNastanka, #kljucneReči
metapodaci za fotografiju: #autor, #datumIVremeNastanka, #mestoFotografisanja, #podešavanjeAparata, #objektiPrikazaniNaFotografiji

3. Kakve sve veze između dokumenta i metapodataka mogu postojati?
Aktivna veza: stanje u kome deo sadržaja jednog dokumenta biva preuzet ili na neki drugi način zavisi od sadržaja drugog dokumenta. Izmenom drugog dokumenta menja se i prvi.
Šta ako se promeni drugi dokument? Da li promena treba da bude vidljiva i u prvom? - upravljanje verzijama dokumenta.

4. Koje su faze životnog ciklusa dokumenta?
Inicijalizacija, priprema, uspostavljanje, korišćenje, revizija, arhiviranje, uklanjanje

5. Objasniti životnu fazu dokumenta korišćenje?
Dokumenti su sa metapodaci dostupni za korišenje. Metapodaci se koriste za pretraživanje i informisanje o dokumentima i njihovim verzijama. U metapodatke se mogu dodati komentari korisnika o korišćenju dokumenta. Distribucija dostavljanje verzija korisnicima na kontrolisani način: automatsko slanje, obaveštenje korisnika o dostupnom dokumentu i lokaciji. Metapodaci vezani za distribuciju: distribucione liste, id primalaca, uloge primalaca u poslovnom procesu, specifikacije formata distribucije, specifikacije formata u kojima je dokument dostupan.

6. Objasniti životnu fazu dokumenta arhiviranje?
Premeštanje dokumenata (verzija, metapodaci) u kompaktniju nepromenljivu formu. Mora da ispini ugovorne obaveze (rok čuvanja). Kontrolisani pristup arhivi. Mogućnost reprodukcije dokumenata. Nemogućnost izmena. Arhiva je baza znanja; potrebni mehanizmi za pretrazivanje. Stabilni, nepromenljivi formati podataka.

7. Objasniti pojmove upravljanje verzijama dokumenata, sekvencijalno i konkurento važenje verzija?
Ako dokument ima više verzija mora imati podršku za upravljanje verzijama. Za svaku verziju postoji period formiranja i period važenja. Važenje verzija se može organizovati sekvencijalno i konkurentno.
Sekvencijalno važenje verzija podrazumeva da je poslednja verzija dokumenta jedina važeća.
Konkurentno važenje verzija podrazumeva da više različitih verzija može biti operativno u jednom trenutku.

8. Koja je osnovna namena sistema za upravljanje dokumentima?
Sistem namenjen praćenju i skladištenju digitalnih dokumenata.

9. Koje su funkcije sistema za upravljanje dokumentima?
Skladištenje dokumenata, katalogizacija, pretraživanje, zaštita podataka, oporavak od katastrofe, arhiviranje, distribucija, upravljanje poslovnim procesima.

10. Opisati Dublin Core format metapodataka.
Mogu se koristiti različiti formati, obično se koristi XML gramatika. DC definiše XML namespace sa elementima koji odgovaraju DC elementima. Ove elemente možemo koristiti u okviru sopstvene XML gramatike ili nekog opšteg standarda.

11. Šta su protokoli za razmenu podataka?
Razmena podataka između sistema putem određenih protokola omogućava jednu vrstu interoperabilnosti sistema.

12. Koje su osnovne karakteristike OAI-PMH protokola?
HTTP baziran, 6 vrsta zahteva, mogućnost definisanje skupova, iterativno preuzimanje.

13. Šta su protokoli za udaljeno pretraživanje?
Omogućavaju da se sa druge mašine postavi upit. Mašina sa koje se postavlja upit i mašina sa koje se dobija upit nisu isti.

14. Koje su osnovne karakteristike Z39.50 i SRU protokola?
XML, SOAP, HTTP, URI, CQI za oba, i za Z39.50 binarni protokoli. SRU je naslednik Z39.50.

15. Čime se bavi oblast pronalaženja informacija (information retrieval)?
Pronalaženje informacija (information retriveval)
reprezentacija, skladištenje, organizacija i pristup informacijama. Pronalaženje materijala (dokumenata) nestruktuirane prirode (text) koji zadovoljava potrebe za infromacijama u okviru velike kolekcije,
16. Koja je razlika između pronalaženja podataka (data retrieval) i pronalaženja informacija (information retrieval)?
Data retrieval – pronalaženje podataka koji zadovoljavaju precizno definisan kriterijum.
Information retrieval – korisnika interesuju informacije o nekoj temi a ne podaci koji zadovoljavaju upit.
17. Kako se razvijala oblast pronalaženja informacija?
*****Pre 4000 godina, *****sadržaji u knjigama, *****indeks pojmova u knjigama, *****indeks na nivou biblioteke knjiga, *****upotreba računara, *****digitalno doba. IR danas – izučava se na većini fakulteta koji se bave računarskim. Google je jedna od najmoćnijih kompanija, osnovna delatnost je IR.
18. Kakve arhitekture mogu imati sistemi za pretragu?
Centralizovani i distribuirani IR sistemi.
19. Koje vrste sadržaja mogu biti pretraživane putem sistema za pretragu?
Po sadržajima u kolekciji – pretraga teksutalnog sadržaja, pretraga linkovanih tekstualnih sadržaja (pretraga veba), pretraga multimedijalnih sadržaja: slika, zvuk, video... , pretraga ostalih vrsta sadržaja..
20. Koji modeli za pretraživanje se koriste u sistemima za pretragu?
Klasični modeli: bulov, vektorski, probabilistički
Alternativni modeli: prošireni bulov, Fuzzy, model neuronske mreže, jezički model
21. Koja je razlika između terma i tokena?
Term - „Normalizovana“ reč (padež, morfologija, itd.); klasa ekvivalencije reči
Token – Instanca reči ili terma koji se pojavljuje u dokumentu.
22. Šta je tokenizacija i koji problemi postoje u ovoj fazi pretprocesiranja?
Tokenizacija – ustanoviti listu tokena, svaki token je kandidat za stavku u listi pojava, ali zavisno od pravila pretprocesiranja neki tokeni će se pojaviti u listi a neki ne.
Problemi: kako procesirati: datume, brojeve, Kinesko pismo (nema whitespace), složenice (Nemački, Finski..)...
23. Zašto se vrši „normalizacija“ reči?
Potrebno je „Normalizovati“ termove u indeksnom tekstu i u upitima u isti oblik. Pojednostavljuje se računanje kosinusa.
24. Šta je to steming?
Gramatičko nepravilno osecanje reči, sa ciljem da se spoje.
25. Šta je to lematizacija?
Pravilno svođenje na ispravan gramatički koren reči.
26. Objasniti Bulov model pretraživanja?
Zasnovan na teoriji skupa i bulovoj algebri. Posmatrani pojam se nalazi ili ne nalazi u dokumentu. Nema rangiranja. Nema parcijalnog poklapanja upita i dokumenta. Konjukcija 3 terma: jednako se posmatra dokument koji nema ni jedan term i dokument koji ima dva terma.
27. Šta je to invertovani indeks i kako se kreira?
Uzima se dokument pa se iz njega uzimaju tokeni termovi da se za svaki od njih vidi gde se pojavljuje. Struktura za pretragu ima rečnik termova za njega listu pojava.
28. Objasniti procesiranje upita kod Bulovog modela?
1. pretprocesiranje upita nakon čega se dobije konjuktivni upit dva terma (ne tokena): nesto AND nesto
2. pronalaženje prvog terma u rečniku, **3.** učitavanje liste pojave prvog terma iz fajla sa pojavama
4 i 5 sve isto samo za drugi term, **6.** izračunavanje preseka ove dve liste pojava; **7.** veraćanje rezultata korisnkiku (dokumenti koji se nalaze u preseku dve liste)
29. Šta su pointeri za preskakanje?
Omogućavaju preskakanje pojava koje svakako neće biti u rezultatu. Izračunavanje preseka može da se ubrza na taj način.
30. Šta se može koristiti ako je potrebno podržati upite fraze?
Dvorečni indeks, duži upiti fraze, proširene dvoreči, pozicioni indeksi.
31. Šta je to dvorečni indeks?
Indeksira svaki susedni par reči u tekstu kao frazu. Može da odgovori na duže upite fraze, ali nakon procesiranja upita koji je izražen pomoću AND operatora. Svaki od parova se tretira kao term u rečniku lako je odgovoriti na dovrečne upite.

32. Šta je to pozicioni indeks?
Pozicioni indeksi su dobra zamena za dovrečne indekse. Omogućavaju odgovore na upite fraze proizvoljne dužine.
33. Objasniti Vektorski model pretraživanja?
Upit ima težinske faktore, ima rangiranje, ima parcijalnog poklapanja upita i dokumenta. U odnosu na bulov model uvodi se stepen slaganja upita i dokumenta vrednost koja se vraća je veća ili jednaka nuli ali nije celobrojna.
34. Šta je ocena relevantnosti?
Ocena je mera koliko se dokument i upit poklapaju. Ako se term ne pojavljuje u dokumentu ocena je 0, što ima više pojava terma u dokumentu ocena je veća.
35. Šta je frekvencija terma?
Frekvencija terma $\mathbf{tf}_{t,d}$ terma t u dokumentu d deniše se kao broj pojavljivanja t u d. Želimo da koristimo tf kada računamo upit/dokument ocene.
36. Šta je frekvencija dokumenta?
Retki termovi su informativniji od čestih. Frekvencija dokumenata je broj dokumenata u kolekciji u kojima se pojavljuje dati term.
37. Šta je tf-idf?
Jedan od najpoznatijih težina. Tf-idf težina terma je proizvod njegove tf težine i njegove idf težine
$W_{t,d} = (1+\log tf_{t,d})* \log(N/df_t)$. Zavistan je od terma i dokumenta.
38. Objasniti kreiranje težinske matrice?
Na pocetku se prvo kreira brojčana matrica i na osnovu tog broja se računa tf
onda može da se radi računanje idf, normalizacija vektora treba da bude jedinična.
39. Koje su razlike između Bulovog i Vektorskog modela pretraživanja?
Bulov model ne omogućava rangiranje rezultata tj. nema parcijanlog parcijalnog poklapanja upita i dokumenta, vektorski model je zasnovan na vektorima u n-dimenzionalnom prostoru pri čemu je n jako velik broj.
40. Da li se relevantnost odgovora meri u odnosu na informacionu potrebu ili upit?
Zadovoljstvo korisnika se može meriti samo prema relevantnosti u odnosu na informacionu potrebu a ne upite.
41. Šta je preciznost (eng. Precision)?
Preciznost P je deo pronađenih dokumenata koji su relevantni u listi pronađenih dokumenata.
42. Šta je povrat (eng. Recall)?
Povrat R je deo pronađenih relevantnih dokumenata u svim relevantnim dokumentima koji postoje u kolekciji.
43. Šta je F mera i zašto je ona relevantnija od korišćenja preciznosti i povrata?
F mera omogućava da se meri kompromis između preciznosti i povrata. Nije važno koliko je dokument relevantan i na kojoj je poziciji u listi rezultata.
44. Kako se može vršiti evaluacija performansi sistema za pretraživanje?
Može sa benchmarkom a postoje i mere web pretraživača.
45. Šta je kapa mera?
Kapa mera je koliko se međusobno ocenjivači slažu, dizajnirana za kategorične ocene.
46. NEĆE BITI!!
47. Šta je to Lucene?
Lucene je javno dostupna biblioteka pisana u javi namenjena pretraživanju teksta.
48. Šta predstavljaju klase Document i Field u Lucene biblioteci?
Klasa document predstavlja jedan dokument koji se indexira, a field predstavlja jedno polje u tom dokumentu. Dokument je skup filedova.
49. Navesti osnovne karakteristike upitnog jezika Lucene-a?
Naziv terma i fileda, može da se kombinuje upiti u bulov upitu, nešto mora biti nešto ne sme, postoji fuzzy, prefix, range i ostali.
50. Kako se implementira analiza (procesiranje) teksta pomoću Lucene-a?
Tokenizatori – rastavlja sadržaj polja na tokene
Token filter – lowercase, stopFilter...
Analizatori – kombinovanje prethodna dva, StandradAnalyzer,SerbianAnalyzer...
...